# Cambridge Scientific Center

36, Y11
October 1966

IBM
Data Processing Division

## Applications of Time-Shared Computers in a Statistics Curriculum

36.Y11

# Applications of Time-Shared Computers in a Statistics Curriculum

M. Schatzoff

IBM Cambridge Scientific Center Report

International Business Machines Corporation
Cambridge Scientific Center
Cambridge, Massachusetts

October, 1966

36.Y11
October, 1966
Scientific Center Report
Limited Distribution

# APPLICATIONS OF TIME-SHARED COMPUTERS IN A STATISTICS CURRICULUM

M. Schatzoff

International Business Machines Corporation
Cambridge Scientific Center
Cambridge, Massachusetts

Abstract

This paper describes the application of remote console computing in a graduate statistics seminar entitled "Machine Aided Statistical Modeling", which was offered at Harvard University by the author during the spring semester 1965-1966. Three different computing systems are discussed: COMB (Console Oriented Model Building), COSMOS (Console Oriented Statistical Matrix Operator System) and the Culler On-Line Computer. The first two systems, designed by the author, operate under the M.I.T. Compatible Time Sharing System. They are directed toward analysis of residual procedures and general linear hypothesis calculations, respectively. The third system, physically located at the University of California at Santa Barbara, is a general purpose on-line computing system featuring a small storage scope.

Applications of all three systems for classroom demonstrations and student exercises are discussed and illustrated.

Index Terms for the IBM Subject Index

Teaching
Statistics
Computations
Time-Sharing
05-Computer Application
16-Mathematics

## LIMITED DISTRIBUTION NOTICE

This report has been submitted for publication elsewhere and has been issued as a Technical Report for early dissemination of its contents. As a courtesy to the intended publisher, it should not be widely distributed until after the date of outside publication.

TABLE OF CONTENTS


| | | |
|---|---|---|
| I. | INTRODUCTION . . . . . | 1 |
| II. | RESIDUAL ANALYSIS . . . . . | 4 |
| III. | THE GENERAL LINEAR MODEL . . . . . | 9 |
| IV. | THE CULLER SYSTEM . . . . . | 17 |
| V. | SUMMARY . . . . . | 19 |
| | ACKNOWLEDGMENTS . . . . . | 20 |
| | REFERENCES . . . . . | 21 |
| | APPENDIX A: Example of Console Session Using COMB . . . . . | A.0 |
| | APPENDIX B: Example of Console Session Using COSMOS . . . . . | B.0 |

# I. INTRODUCTION

The past decade has witnessed the widespread and ever growing use of electronic computers by universities, government, industry and private research facilities. Applications ranging from routine data processing to highly sophisticated modeling, and encompassing virtually every scientific discipline and type of business activity, have been encountered. However, despite the fact that advances in computer technology have been rapid and impressive, they have not always been accompanied by corresponding strides in many areas of computer application. With rare exception, the entire area of statistical computer application, which has been characterized mainly by a proliferation of "canned" programs, has been rather routine and unimaginative. Moreover, because of the wide availability of such programs, the teaching of statistical computation and data analysis has been de-emphasized instead of being re-directed to take advantage of the computer.

We are just now at the threshold of a new era in computing, entering the age of the time-shared computer. Briefly, a time-sharing system allows many users to access a large central computer simultaneously from remote terminals such as typewriters, keyboards or scopes. The individual user can typically enter data and instructions, compile, edit, load or execute programs, and obtain responses from the computer so rapidly that in effect he may feel as if he is the sole user of the computer. This nearly instantaneous communication facility provided between the individual and the com-

puter will doubtlessly lead to countless hitherto unimagined applications of great potential value. From the standpoint of the statistician engaged in data analysis, a remote console may be viewed as a powerful computational tool, supplanting, and far exceeding the capabilities of conventional devices such as desk calculators, graph paper and tables of functions. To assure widespread use of computer consoles in this fashion, it is essential that software systems designed for such purposes be easy to learn and easy to use by the computer novice. In effect, the non-programmer statistician should be readily able to use the console as a computational tool for carrying out, on large bodies of data, virtually any calculations that he might be able to perform on small bodies of data using tools such as desk calculators, graph paper and tables of functions.

One of the most important features of any digital computer is its ability to make decisions during the execution of a program by means of programmed conditional branching instructions; a corresponding innovation of great importance in a time-shared computing environment is the facility of introducing human decision making within the program. That is, because of the rapid man-machine communication facility afforded by a time-shared computing system, the user can direct the computer to produce a sequence of complex computations, receive the results of such computations (in either numerical or graphical form) immediately, and then decide, based on examination of such results, what computations he would like the computer to carry out next. Exploitation of problem areas characterized by the requirement of this type of rapid man-machine interaction will produce valuable problem solving approaches not previously available.

Having attempted to motivate the ensuing discussion, we shall proceed

to describe the application of remote console computing in a graduate statistics seminar offered at Harvard University by the author in the spring semester 1965-1966. A brief description of the course, taken from the catalogue of the Harvard Graduate School of Arts and Sciences, follows: "Statistics 285. Machine Aided Statistical Modeling.

Application of time-shared computing to the construction and testing of statistical models. Topics will include methods for analyzing residuals from a least squares fitting and assessing the validity of underlying assumptions. Students will use remote consoles to generate and analyze data from a variety of mathematical models.

Prerequisite:

Statistics 139 (Analysis of Variance). Knowledge of programming is not required."

Lectures were held in a room at the Harvard Computation Center equipped with a closed circuit television receiver which was used to display console operation televised from a nearby machine room. Two remotely accessed computer systems, the MIT Compatible Time-Sharing System (based on the IBM 7094) and the Culler On-Line Computer (based on the TRW 400) were used for classroom demonstrations and student projects. The principal sections of this paper deal with statistical applications of the first of these two systems in the aforementioned course, as demonstrated on-line by the author at the 1966 Joint Statistical Meetings in Los Angeles, August 16, 1966 (Schatzoff, (1966)).

In Section 2, we describe the operation of COMB (Console Oriented Model Building), a statistical computing system designed to implement and study the analysis of residuals procedures of Anscombe and Tukey (1963). We then proceed in Section 3, to discuss a second statistical system, COSMOS (Console

Oriented Statistical Matrix Operator System), which employs the basic operators defined by Beaton (1964) to tackle problems associated with the general linear model. In Section 4, we indicate briefly some of the applications of the Culler system, which is used primarily to manipulate and display functions visually on an electronic storage scope, and conclude in Section 5 with some remarks relative to future implications.

## II. RESIDUAL ANALYSIS

There has been considerable interest, in recent years, in methods for examining residuals from a least squares fitting, with a view to assessing the validity of the usual model assumptions. An excellent summary of a number of such procedures is provided by Anscombe and Tukey (1963). Use of a time-shared computer to implement the Anscombe-Tukey procedures has been described previously by the author (Schatzoff (1965); in this section, we shall describe a much later version of this computational approach, by illustrating the use of the COMB system from a typewriter console. The name COMB is meant to imply also the ability to comb a set of data in a number of directions in order to ascertain what the data has to tell us.

The COMB system was implemented initially under the MIT Compatible Time-Sharing System. In effect, when the user types commands at a typewriter console, the full computational power of an IBM 7094 computer is made available to him. In following the examples provided in the appendices, the reader should bear in mind that the typewriter console provides a two way communication channel. Messages typed by the user appear in lower case type, while output from the computer is in upper case. Thus, it is easy for the reader to interpret each of the examples as a two way dialogue between the statistician and the computer.

Referring to Appendix A, the user initiates his session by typing the command "r comb", which is interpreted by the computer as meaning

"resume operation of a program named comb". The computer replies with the message "W 1020.7" meaning "Wait until I find your program called comb, (which is stored permanently on a magnetic disc file) and activate it (i.e.-bring it into core storage and begin execution). The time of day is now 1020.7." The computer then types the message "READY" a few seconds later, indicating that the COMB program has been activated and is ready for the user to proceed.

It should be pointed out at this time that COMB is tailored to the residual anaylsis procedures for a two way fixed effects analysis of variance model. Since the program was intended primarily as a pedagogical device, it provides facilities for generating data from such a model, with provisions whereby the user can specify any of a number of types of departures from the usual normal theory assumptions. The types of departures which may be specified are precisely those which the Anscombe-Tukey procedures are supposed to detect, so that the system provides a means for investigating the behavior of these procedures.

Returning to the example, the user types the command "start" which tells the computer to start operation of the program by allowing him either to enter or generate a set of data. The ensuing question and answer session is fairly self-explanatory. Depending on the responses given at each point by the user, he may type in a set of data, use the data saved from his previous session at the console, or have the computer generate a set of data from a mathematical model specified by him. If the latter option is selected, the user must indicate the size of the experiment; additionally, he must either request the computer to generate numerical values of the parameters of the model, or as was the case in this example, type in these values himself. The model is of the form:

(1) $$Y_{ijk} = \mu + \alpha_i + \beta_j + G\alpha_i\beta_j + e_{ijk}$$
where the $e_{ijk}$ are independent $N(0,\sigma^2)$.

In the example, an asterisk denotes multiplication, while a quote symbol denotes electronic erasure of the previous character. Thus, the last q characters typed on a line may be erased from the computer's memory by typing q quote marks. An entire line may be erased by means of the question mark. As evidenced by the series of questions, the user may specify either an additive model (G=0) or an interactive model ($G\neq 0$), and may perturb the model by:

1. Sampling each error term $e_{ijk}$ from a normal distribution whose variance depends on the level $\mu_{ijk} = E(Y_{ijk})$. The nature of the dependence is $\sigma^2_{ijk} = a \exp(b\,\mu_{ijk})$, and if this option is selected, the user receives instructions for selecting a and b.
2. Sampling the error terms from a scale contaminated normal distribution $p\,N(0,\sigma^2) + (1-p)\,N(0,k^2\sigma^2)$, where the user may specify p and k.
3. Specifying a fixed number of outliers of the form
$$Y_{ijk} = \mu + \alpha_i + \beta_j + G\alpha_i\beta_j + e_{ijk} + K \quad .$$
The locations of the outliers (i.e. - the indices i, j and k) are selected at random by the computer.

When the data generation has been completed, the computer types the message "READY" signifying that the analyst may proceed to issue commands. A list of available commands (or codes), together with their definitions, may be obtained by typing the command "list", as shown in the next section of the example. As indicated by this list of commands, one can print the data, fitted values, residuals and estimates of the mean and main effects, perform any of a number of statistical analyses and tests, obtain specialized

plots, make transformations on the observations or change the model itself. Further, he can at any time retrieve the original data, process a new set of data or use the computer as a desk calculator. Finally, the user can, by typing the command "enlite" be "enlightenend" regarding the model from which the data were generated. In response to this command, the computer will type the values of all the parameters of the model, indicate the nature of the dependence of variance upon level of response if applicable, and identify the locations and values of any contaminated observations or outliers. Thus, the instructor can use the system to generate the data to be analyzed by students, and they in turn can be "enlightened", at the conclusion of their analyses, as to the "true state of nature".

The user may issue one or more commands at a time, in any sequence whatever. Thus, turning to Section A.3, we see that the four commands "data res fit est" may be typed on a single line, and the computer will print the data matrix, residuals, fitted values and estimates of the mean and main effects before indicating that it is "READY" for further instructions. In Section A.4, the command "shape" calculates estimates of the third and fourth standardized moments (G1 and G2) together with their standard errors and corresponding t values for testing normality, as per Anscombe (1961). This is followed by Tukey's (1949) one degree of freedom test for nonadditivity, where K is the regression coefficient of the residuals on the squared fitted values, F is the F-ratio for the one degree of freedom test, and P is a suggested power transformation for removing nonadditivity. It should be noted that the data generation phase of the program permits generation of a nonadditivity term of the form assumed by Tukey, and that the list of commands includes the facility for making power transforma-

tions. Thus, it is easy to learn, on an empirical basis, how well the test and the transformation work. The third command in the sequence, "errdep", produces calculations associated with Anscombe's (1961) test of error variance dependent on level. Here, H is a linear regression coefficient of the squared residuals on the fitted values ($Y_{ij}$), T is an approximate t-statistic for testing the null hypothesis $H_o$: H=0, and P is a suggested power transformation for removing the type of dependence assumed by Anscombe. Note again that the program permits generation of data from just such a model. Finally, an analysis of variance table can be produced by typing the command "anova ", and the observed significance level corresponding to any observed F-ratio can be computed with the command "fdist". The example is concluded in Section A.5 with a full normal plot (Tukey (1962)) of the residuals, and "enlightenment". The cell index corresponding to each residual is indicated at the left hand margin of the full normal plot, which is scaled at the top in standard deviation units of the quantities being plotted. Thus, the square root of the mean square for error is indicated on the top of the plot as S=1.204. Directly below it is the estimate of $\sigma$ obtained by averaging the points on the full normal plot. The middle one-third of the points are suppressed from the plot, as suggested by Tukey, since they have a large variance. Furthemore, most types of departures from the standard assumptions, such as skewness, heavy tails, and outliers, will be evidenced in the tails. Samples from a normal distribution should plot in a vertical straight line, centered at the standard deviation of the distribution.

In the example, we have not used all of the facilities provided in the program. However, it should be clear that one could continue to perform tests, make transformations, and change model assumptions in any sequence whatever. The various tests and plots are designed to reveal particular aspects of the data, and the analyst can ponder the results at each

stage and decide what to do next. The command structure is extremely simple, so that the user does not have to know anything at all about computers or computer programming. In the course, lecture sessions covered the theory underlying the various residual analysis procedures incorporated in the program, and the closed circuit television facility was used to demonstrate the application of these procedures. Students were provided access to the consoles so that they could experiment with these and other techniques, and in effect gain a great deal of data analysis experience in a very short period of time. On occasion, data was generated by the instructor and given to the students to analyze, without providing the "enlightenment" facility. Classroom discussion of the analyses and comparison of results usually proved interesting and illuminating.

## III. THE GENERAL LINEAR MODEL

Statistical analyses associated with the general linear model typically involve but a few basic types of matrix operations, such as linear transformations, solution of simultaneous linear equations, successive orthogonalization of variables and eigenvalue - eigenvector analysis. Beaton (1964) in a doctoral dissertation submitted to the Harvard Graduate School of Education, illustrated the use of six basic matrix operations to carry out standard calculations required in correlation analysis, regression analysis, analysis of variance and covariance, and the usual collection of multivariate procedures such as multivariate regression and analysis of variance, discriminant analysis, principal component analysis and canonical correlation analysis. It is pedagogically attractive for students to perform these kinds of analyses in terms of matrix operations since they would have learned the underlying theory in terms of matrices. Such an approach obviates the necessity of horrendous desk calculations yet preserves the

spirit of requiring the user to understand what he is doing, rather than permitting him to use canned programs, which may not possess the flexibility to provide what he really wants.

In this section, we illustrate the operation of an on-line statistical computing system called COSMOS (Console Oriented Statistical Matrix Operator System), which includes all of the Beaton operators, and has a simple command structure. Data may be entered from an IBM 1050 typewriter console, and commands may be executed one at a time, by typing one command per line, or sequentially, by typing several commands on a line. In general, the matrix result of any operation may be used as an input matrix to the subsequent operation, and intermediate results may be printed or saved under matrix names provided by the user for later reference. Commands may be executed in any sequence specified by the user, so that the system provides great flexibility in allowing the user to tailor the analysis to the particular situation. Finally, there is a "macro" facility which enables the user to create new commands consisting of specified sequences of system commands and other "macro" commands. Thus, the basic commands in the system can be used to create higher level commands and "canned" programs in a very simple manner.

A command in the COSMOS system consists of the name of an operator, followed by the name of a matrix (or matrices) upon which the operator is to act, and a list (or the name of a list) which refers to the range of the operation within the matrix. Omission of the list implies operation on all rows and columns of the matrix. For example, the command

(3.1) prm data .

means print the matrix called data, whereas the command

(3.2) prm data 1 2 , 2 4 8 .

means print the matrix defined by the intersection of rows 1 and 2 with columns 2 4 and 8 of the matrix called data. The same result could have been achieved by typing the command sequence

(3.3) stl dick 1 2 3 · stl jane 2 4 8· prm data dick , jane .

where the command stl dick 1 2 . means store a list called dick containing the elements 1 2 .

Since it is frequently desirable to use the result of an operation as input to the subsequent operation, omission of a matrix name in the command implies that the previous matrix result, referred to by the COSMOS system as WKM1 (work matrix 1), will be used as the input matrix. For example, to print the result of the previous matrix operation, one simply types the command

(3.4) prm ·

Having introduced these few preliminaries concerning the system usage conventions, we will proceed to define three of the six basic matrix operators and illustrate their use in the system to carry out stepwise correlation and regression analyses.

Suppose we have a data matrix $D_{nxp}$ consisting of n observations on p variables. The command

(3.5) scp d listr , listc ·

means calculate the sum of cross products matrix D*' D*, where the submatrix D* is defined by the intersection of the rows of d specified by listr and the columns specified by listc. If the lists are omitted, the command computes D'D. (It should be noted that the matrix "D" is typed "d" since the user is restricted to lower case.) Suppose now that we have a square matrix $A_{pxp}$, which might have resulted from having executed an SCP command. The command

(3.6) swp a k .

with k a single element list, means sweep the matrix A on the $k^{th}$ pivotal element. It will produce a matrix B defined by

(3.7)
$$B_{kk} = {}^{1}/A_{kk}$$
$$B_{kj} = A_{kj}/A_{kk} \qquad j \neq k$$
$$B_{ik} = -A_{ik}/A_{kk} \qquad i \neq k$$
$$B_{ij} = A_{ij} - A_{ik} A_{kj}/A_{kk} \qquad i,j \neq k .$$

The use of swp with a multiple element list has the effect of sweeping the matrix on the pivotal element specified by the first element of the list, then sweeping the resulting matrix on the pivotal element specified by the second element of the list etc. Thus, the command

(3.8) swp a r s ... t .

is equivalent to the command sequence

(3.9) swp a r . swp s . ... swp t .

It is easy to check from (3.7) that the swp operation is commutative and reversible. That is, application of the sweep operator with a given list is equivalent to sweeping with any permutation of the elements of that list, and sweeping twice on a given pivotal element is equivalent to not having swept on that element at all. Finally if we consider the matrix A partitioned along the first m rows and columns as per

(3.10)
$$A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right) \begin{array}{l} 1 \\ \vdots \\ m \\ \hline m+1 \\ \vdots \\ p \end{array}$$

and issue the command

(3.11) swp a 1 2 ... m .

we obtain the resulting matrix

$$(3.12)\qquad \left(\begin{array}{c|c} A_{11}^{-1} & -A_{11}^{-1}A_{12} \\ \hline -A_{21}A_{11}^{-1} & A_{22}-A_{21}A_{11}^{-1}A_{12} \end{array}\right) \begin{array}{c} 1 \\ 2 \\ \vdots \\ m \\ \hline m+1 \\ \vdots \\ p \end{array}$$

The third operator which we shall employ is called standardize, and abbreviated std. The command

(3.13) std a r s ... t .

transforms the matrix A into a matrix B with elements defined by

$$(3.14)\qquad B_{ij} = A_{ij} / (A_{ii} A_{jj})^{1/2}$$

for all i, j belonging to the list.

Appendix B contains an example of a console session using the COSMOS system and the same set of data used by Beaton (1964), who borrowed the example from Walker and Lev (1953). The data consists of observations on six variables for each of 98 students. The six variables are:

$X_1$ Reading Score

$X_2$ Artificial Language Score

$X_3$ Arithmetic Score

$X_4$ Mid-Term Test Score

$X_5$ Final Examination Score

$X_6$ Semester Grade .

Additionally, a dummy variable having the value 1 for all observations was appended to the data matrix for purposes of convenience, which will become apparent shortly. We start the analysis at the point where the data has already been entered into the computer by means of the command stm (store a matrix) and filed permanently under the name "data".

Referring to Appendix B, the first sequence of commands computes and

prints the sum of cross products matrix for the six variables and the dummy variable. This matrix is of the form

(3.15)
$$\begin{pmatrix} \sum x_1^2 & \sum x_1x_2 & \cdots & \sum x_1x_6 & \sum x_1 \\ \sum x_1x_2 & \sum x_2^2 & \cdots & \sum x_2x_6 & \sum x_2 \\ \vdots & \vdots & & \vdots & \vdots \\ \sum x_1x_6 & \sum x_2x_6 & \cdots & \sum x_6^2 & \sum x_6 \\ \sum x_1 & \sum x_2 & \cdots & \sum x_6 & n \end{pmatrix}$$

where the summations are taken over the n observations, and in the example, $n = 98$.

The next sequence of commands sweeps the sum of cross products matrix on the $7^{th}$ pivotal element, prints the result, and stores three lists representing all the variables $(X_1, \ldots, X_6)$, the independent variables $(X_1, \ldots, X_3)$ and the dependent variables $(X_4, \ldots, X_6)$. The function of the dummy variable $(X_7)$ is laid bare by the command swp 7 · which transforms the sum of cross products matrix (3-15) to a mean-centered cross products matrix. Thus, one can check directly from the definition of swp (3.7) that the (i, j) element of the resulting matrix will be given by $\sum x_i x_j - (\sum x_i)(\sum x_j)/n$.

The third sequence of commands specifies that the mean-centered cross products matrix resulting from the previous sequence of commands is to be saved under the name "cov", and that all the variables are to be "standardized" and printed. As the reader may check from (3.14), the result of this operation is the matrix of simple correlation coefficients for the six variables.

Next, we see from (3.12) that sweeping the mean-centered cross products

matrix ("cov") on the independent variables (1 2 3) and printing the seventh row and column first, results in the matrix

(3.16)

$$\left(\begin{array}{c|c|c} & & 7 \\ & & 1 \\ (X'X)^{-1} & (X'X)^{-1}X'Y & 2 \\ & & 3 \\ \hline & & 4 \\ -Y'X(X'X)^{-1} & Y'Y - Y'X(X'X)^{-1}X'Y & 5 \\ & & 6 \end{array}\right)$$

where $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{13} \\ 1 & x_{21} & \cdots & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n3} \end{pmatrix}$ $Y = \begin{pmatrix} x_{14} & \cdots & x_{16} \\ x_{24} & \cdots & x_{26} \\ \vdots & \vdots & \vdots \\ x_{n4} & \cdots & x_{n6} \end{pmatrix}$

It is clear from the least squares theory that the columns of the upper right hand sub-matrix contain the least squares estimates of the regression coefficients of $X_4$, $X_5$ and $X_6$ on $X_1$, $X_2$, and $X_3$, while the lower right hand submatrix is (n-4) $S_{2.1}$, where $S_{2.1}$ is the residual covariance matrix of the dependent variables ($X_4$, $X_5$ and $X_6$) on the independent variables ($X_1$, $X_2$ and $X_3$). Standardizing the latter submatrix therefore produces the matrix of partial correlation coefficients of the dependent variables, after removing the effects of the independent variables.

The next section of the example, starting with the command swp cov 1 . , shows the steps required to obtain all seven possible regressions of the dependent variables on the independent variables in exactly seven sweeps. It can in fact be readily shown that all $2^k - 1$ regressions of a set of dependent variables on a set of k independent variables can always be obtained with $2^k - 1$ sweeps of the mean-centered sum of cross-products matrix.

Next, the use of the macro instruction to create a "canned" regression program is illustrated. The sequence of commands required for calculating and printing correlation coefficients, least squares estimates of regression coefficients and residual sums of cross-products, and partial correlation coefficients, are typed after the word "macro" and the name "regr" is assigned to the sequence. The commands are not carried out at this time, but rather, the command "regr", which consists of the given sequence of commands, is added to the list of COSMOS commands. Now, by merely typing the command "regr" the initial regression and partial correlation analysis can be carried out. Any number of such macros can be created by the user, and a macro command such as "regr" can itself be part of another macro. In this manner, the user can create quite intricate programs with relative ease. On typing the command "quit", the computer types the user dictionary, which contains the names of all matrices, lists and macros currently defined in the system. The user is then provided with instructions for saving such information if he so desires, and the computer types "R" (meaning "ready" for further commands) and provides an indication of the computer time, in seconds, required for operating the program (40.716) and for swapping the program in and out of core storage to accomodate other users (14.550). Finally, the user logs out, physically disconnecting his terminal from the computer.

It should be remembered that we have illustrated but a few of the commands available in the COSMOS system. As indicated earlier, the system includes other commands which greatly facilitate stepwise multivariate analyses. These basic operators, when used in conjunction with operators designed to allow a wide variety of data manipulation and

selection options, transformations, sampling procedures and indexing and branching features, will permit very flexible, tailored analyses of highly complex bodies of data with relative ease. By using the COSMOS system, which is very easy to learn, the student can obtain valuable experience in analyzing real data, just as fast as such data can be provided for his use. Accordingly, one of the aims of the COSMOS system is to provide a data bank consisting of real data from a variety of application fields. It is hoped that widespread availability of COSMOS and appropriate computing facilities will encourage statisticians to analyze data themselves at computer consoles, and to experiment with new techniques of data analysis.

## IV. THE CULLER SYSTEM

The Culler on-line computer, physically located at the University of California at Santa Barbara, is operated from specially designed remote consoles which present the user with a keyboard for entering data and instructions and a small electronic storage scope for displaying instructions, data and functions. Each button on the keyboard performs a specific function such as add, multiply, sin, log, display, sort, etc. There are a number of different levels of operation within the system, so that if the system is in the vector level for example, as specified by pushing the appropriate vector usage level button, the add button signifies addition of vectors, whereas if the system is in a scalar level, one may operate with scalars, including individual components of vectors or matrices. Other levels are provided for operating with real and complex arrays, special functions, and user-defined functions.

Because the system is designed to operate with "one-button pushes", it is very easy to use. For instance, one can construct and display

density and distribution functions quite readily, draw samples from these distributions and display the cumulative sample functions on appropriate scales such as normal probability, full normal, half-normal, or any other desired scale with but a few button pushes. There is also a macro facility provided in the system, so that the user can construct and store subprograms which themselves may be activated by depressing a single button. Subprogram commands may themselves be included as commands in other subprograms so that fairly complex programs can be created and stored in the computer.

Classroom utilization of the Culler system was primarily devoted to the types of operations described above. Density and distribution functions for normal, chi-square and contaminated normal distributions were constructed and displayed on the scope. We could then draw repeated samples from any of these distributions and plot the cumulative sample functions on arithmetic, normal or full normal scales. Outliers could be introduced into the samples, and the cumulative sample functions could be displayed on any of the indicated scales before and after techniques such as Winsorization, trimming, rejection of outliers, or other types of transformations had been employed. Utilization of the Culler system in this manner provided considerable insight into the operation of the indicated techniques as well as lively classroom discussion, which would frequently lead to suggestions which could be implemented spontaneously by pushing a few buttons. Such sessions were generally informative and enjoyable.

As with the MIT Compatible Time Sharing System and the two statistical systems described in sections 2 and 3, students had access to the Culler System, and everyone managed to get a fair amount of hands-on

time on one or the other (or both) of the systems. Each student was required to present a paper to the class, describing a project undertaken on one of the two computers.

## V. SUMMARY

The use of time-shared computers as an aid in teaching techniques of statistical model building and data analysis at Harvard University has been described in some detail. We have been fortunate in having terminals for the two computer systems available at Harvard so that a course of the nature described here could be undertaken. Within the next few years, many universities throughout the country will enjoy the availability of similar computational facilities; it is hoped that this revolutionary type of computing service will lead to a corresponding revolution in the use of computers by statisticians and other scientists engaged in research or teaching.

The course described in this paper was of necessity experimental in nature, and as with any other course offered for the first time, it is bound to undergo many changes the next time it is given. However, I am sure that the undertaking constituted a step in the right direction, and that similar applications of computers will be explored on a wider basis in the coming years. The prudent integration of on-line computing into a variety of statistics courses, through classroom use and homework exercises, would, I believe, constitute an important advance in the teaching and learning of statistical methods.

In assessing the value of the course, it is felt that a number of worthwhile objectives were achieved. First, students were able to better understand the particular methods studied by using these methods to analyze data and observing their behavior under a variety of model as-

sumptions. Experience in data analysis is itself important in the training of any statistician, whether that training be oriented to theory or practice, for one develops insight into data analysis problems very rapidly in the course of working with data. A fringe benefit was that of introducing some of the students to computers for the first time; it is perhaps not unreasonable to require that some sort of computer training should be mandatory for every statistics student. If such training is oriented to statistical applications, it is perhaps more palatable than a general purpose computer course. Finally, it is my hope (and belief) that some of the students will have been sufficiently motivated to develop their own approaches to the intelligent use of computer resources in research, teaching and data analysis, for a revolution in computing is truly upon us, and statisticians should be ready to meet the challenge.

ACKNOWLEDGMENTS:

I wish to express my gratitude to Professor Fredrick Mosteller for allowing me to experiment with his students in the manner indicated, and to Professor Anthony Oettinger for having the foresight to have CTSS and Culler terminals installed at the Harvard Computation Laboratory, along with closed circuit television facilities to enhance their use in the classroom. Dr. William Bossert provided valuable assistance in the use of the Culler System and preparation of classroom demonstrations.

Finally, it is with pleasure that I acknowledge the invaluable contributions of several staff members of the IBM Cambridge Scientific Center in the design and programming of the COMB and COSMOS systems. Foremost among these are Mr. Thomas Burhoe and Dr. Arthur Anger.

REFERENCES:

1. Anscombe, F. J., "Rejection of Outliers", Technometrics, 2 (1960), 123-147.

2. Anscombe, F. J., "Examination of Residuals", Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability" (1961), (University of California Press), 1, 1-36.

3. Anscombe, F. J. and Tukey, J. W., "The Examination and Analysis of Residuals", Technometrics, 5 (1963), 141-160.

4. Beaton, Albert E., "The Use of Special Matrix Operators in Statistical Calculus", Research Bulletin RB-64-51, (Educational Testing Service) Princeton, New Jersey, (1964).

5. Schatzoff, M., "Console Oriented Model Building", Proceedings of the 20th National Conference of the Association for Computing Machinery, (1965), 354-374.

6. Schatzoff, M., "Uses of Computers in a Statistics Curriculum: An On-Line Demonstration", Abstracts Booklet, Summaries of Papers Presented at the Joint Statistical Meetings of the American Statistical Association, Biometric Society (Eastern and Western North American Regions), Institute of Mathematical Statistics (Western Regional Meeting) and Western Farm Economics Association (American Statistical Association), 1966.

7. Tukey, J. W., "One Degree of Freedom for Nonadditivity" Biometrics, 5 (1949), 232-242.

8. Tukey, J. W., "The Future of Data Analysis", Annals of Mathematical Statistics, 33, (1962), 1-67.

9. Walker, Helen M., and Lev, Joseph, Statistical Inference, Henry Holt and Company, New York, 1953.

# APPENDIX A

## Example of Console Session Using COMB

```
r comb
W 1020.7                                A.1
READY
start
DO YOU WISH TO INPUT YOUR OWN DATA, YES OR NO.
no
DO YOU WISH TO USE DATA FROM YOUR LAST SESSION.
no
 INPUT NUMBER OF ROWS, COLUMNS AND REPLICATIONS.
6,6,1
WANT RANDOM GENERATION OF PARAMETERS
no
 TYPE MU AND G, WHERE MU IS THE GRAND MEAN AND THE (I,J) INTERACTION TERM IS
 G*ALPHA(I)*BETA(J).
10,0
 INPUT ROW MAIN EFFECTS.
-2.5,-1.5,-.5,.5,1.5,2.5
 INPUT COLUMN MAIN EFFECTS.
-1,-.4,0,.1,.5,.8
 WANT ERROR VARIANCE TO DEPEND ON LEVEL
no
 INPUT SIGMA SQUARED.
1
 WANT CONTAMINATED ERRORS
yes
 TYPE SIGMA MULTIPLIER AND PROBABILTY.
5,.1
 WANT OUTLIERS
no
READY
```

list



| | |
|---|---|
| LIST | DESCRIPTION OF CODES |
| DATA | PRINT DATA MATRIX |
| EST | PRINT ESTIMATES OF MEAN AND MAIN EFFECTS |
| FIT | PRINT FITTED VALUES |
| RES | PRINT RESIDUALS |
| OUTL | OUTLIER TEST, USER MUST SPECIFY PREMIUM |
| ANOVA | ANALYSIS OF VARIANCE TABLE |
| SHAPE | THIRD AND FOURTH MOMENTS OF RESIDUALS, TOGETHER WITH T |
| NONADD | ONE DEGREE OF FREEDOM TEST FOR NON-ADDITIVITY |
| FDIST | UPPER TAIL PROBABILITY OF F DISTRIBUTION, USER MUST SPECIFY F STATISTIC AND DEGREES OF FREEDOM |
| ERRDEP | DEPENDENCE OF VARIABILITY UPON LEVEL OF RESPONSE |
| FUNOP | FUNOP PLOT |
| PLOTRF | PLOT OF RESIDUALS VS. FITTED VALUES |
| LOG | LOGARITHMIC TRANSFORMATION |
| ASINY | ARCSINE TRANSFORMATION |
| POWER | POWER TRANSFORMATION, USER MUST SPECIFY EXPONENT |
| ADDK | ADD A CONSTANT TO EACH OBSERVATION |
| MPYK | MULTIPLY EACH OBSERVATION BY A CONSTANT |
| CHGY | MODIFY INDIVIDUAL OBSERVATIONS |
| CHGMOD | CHANGE THE MODEL FROM ADDITIVE TO INTERACTIVE OR VICE VERSA |
| RECOUP | RETRIEVE THE ORIGINAL DATA |
| START | PROCESS A NEW SET OF DATA, PREVIOUS DATA CANNOT BE RETRIEVED |
| DESCAL | DESK CALCULATOR |
| CODES | LIST OF CODES |
| QUIT | TERMINATE SESSION |

READY

```
data res fit est
```



DATA MATRIX

| | | | | | |
|---|---|---|---|---|---|
| 6.392 | 6.627 | 6.921 | 9.470 | 7.907 | 7.780 |
| 5.671 | 8.762 | 8.459 | 8.955 | 9.419 | 10.097 |
| 6.826 | 8.509 | 9.644 | 13.235 | 11.333 | 12.899 |
| 8.474 | 9.380 | 10.206 | 11.174 | 7.913 | 11.348 |
| 10.869 | 11.849 | 10.231 | 11.049 | 12.286 | 11.422 |
| 10.623 | 10.673 | 10.470 | 9.914 | 12.696 | 12.976 |

RESIDUALS

| | | | | | |
|---|---|---|---|---|---|
| .524 | -.399 | -.127 | 1.111 | -.077 | -1.032 |
| -1.242 | .692 | .367 | -.448 | .390 | .240 |
| -1.934 | -1.408 | -.295 | 1.985 | .457 | 1.195 |
| .373 | .122 | .926 | .582 | -2.305 | .302 |
| 1.233 | 1.055 | -.584 | -1.078 | .533 | -1.159 |
| 1.046 | -.062 | -.286 | -2.153 | 1.002 | .454 |

FITTED VALUES

| | | | | | |
|---|---|---|---|---|---|
| 5.868 | 7.026 | 7.048 | 8.358 | 7.985 | 8.813 |
| 6.912 | 8.070 | 8.092 | 9.403 | 9.029 | 9.857 |
| 8.759 | 9.917 | 9.939 | 11.250 | 10.876 | 11.704 |
| 8.101 | 9.259 | 9.281 | 10.591 | 10.218 | 11.046 |
| 9.636 | 10.794 | 10.816 | 12.126 | 11.753 | 12.581 |
| 9.577 | 10.735 | 10.757 | 12.067 | 11.694 | 12.522 |

ESTIMATED MEAN
9.7905

ESTIMATED ROW MAIN EFFECTS
-2.2744
-1.2300
.6171
-.0413
1.4938
1.4348

ESTIMATED COLUMN MAIN EFFECTS
-1.6482
-.4904
-.4685
.8421
.4685
1.2965

READY

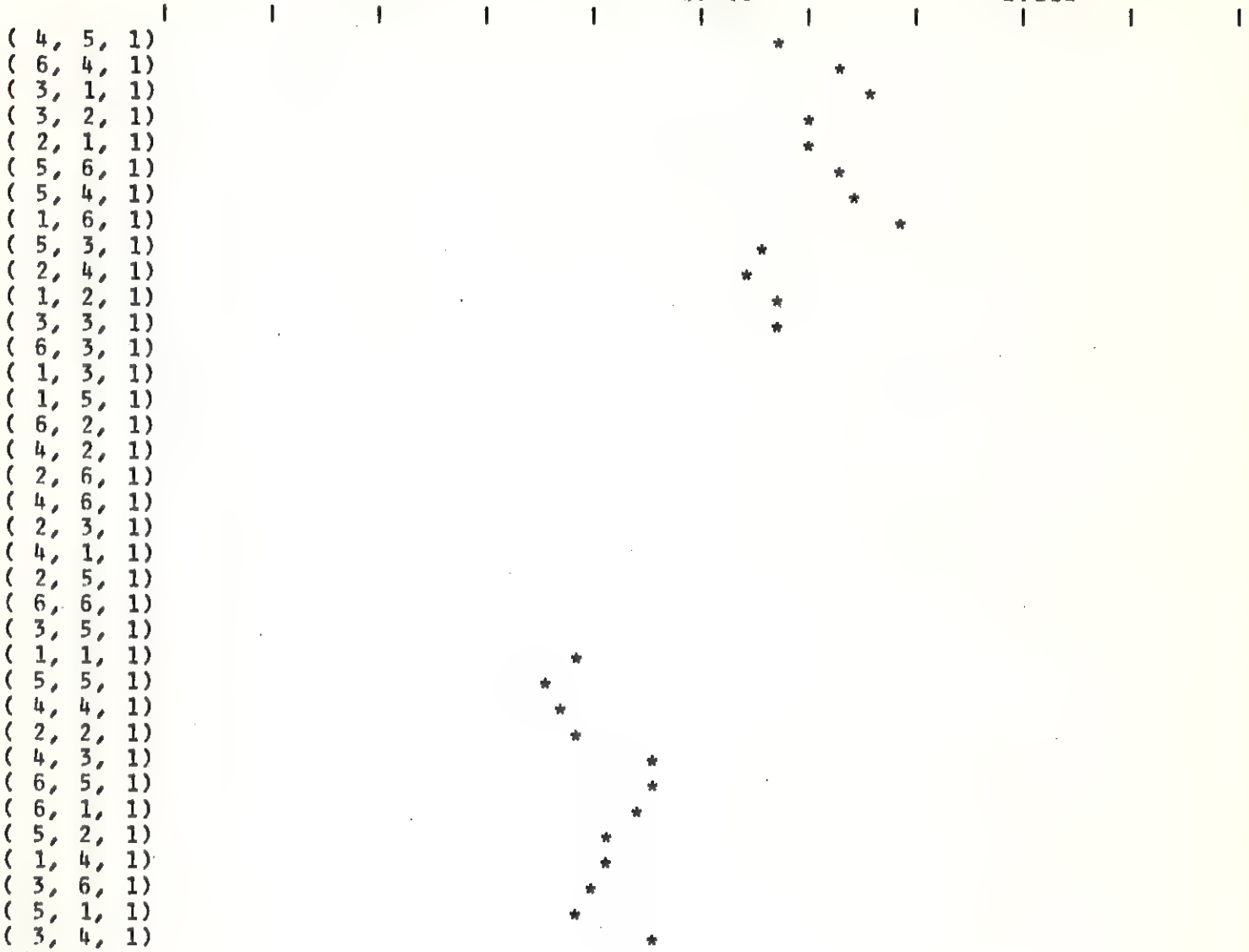
```
shape nonadd errdep anova
 G1 =    -1.0494, G2 =      -.0447
 STANDARD ERROR OF G1 =       .6565
 STANDARD ERROR OF G2 =      1.4682
 T1 =    -1.5984, T2 =      -.0305


K =      -.0376
F =       .2492
P =      1.7371


     H       VAR(H)        T         P

     .2192      .0324     1.2177    -.0731

SOURCE          SS          D.F.     MS           F

ROW           68.1504         5     13.6301      9.3992
COLUMN        34.7158         5      6.9432      4.7879
ERROR         36.2534        25      1.4501
TOTAL        139.1196        35


READY
fdist
 INPUT F, DF1, DF2
4.7879,5,25

    F        DF1   DF2     ALPHA
    4.7879     5    25      .0033


 INPUT F, DF1, DF2
INT. 0
READY
```

funop

A.5

S= 1.204

-.107 1.003 2.112




READY

enlite

| MU | GAMMA |
|---|---|
| 10.000 | 0. |

SIGMA = 1.000

ROW MAIN EFFECTS

| | | | | | |
|---|---|---|---|---|---|
| -2.500 | -1.500 | -.500 | .500 | 1.500 | 2.500 |

COLUMN MAIN EFFECTS

| | | | | | |
|---|---|---|---|---|---|
| -1.000 | -.400 | 0. | .100 | .500 | .800 |

CONTAMINATED ERRORS

| | | | | | |
|---|---|---|---|---|---|
| Y( 1, 4, 1) | HAS BEEN CHANGED | FROM | 6.841 | TO | 9.470 |
| Y( 2, 4, 1) | HAS BEEN CHANGED | FROM | 8.810 | TO | 8.955 |
| Y( 3, 4, 1) | HAS BEEN CHANGED | FROM | 7.766 | TO | 13.235 |
| Y( 3, 6, 1) | HAS BEEN CHANGED | FROM | 9.486 | TO | 12.899 |
| Y( 6, 4, 1) | HAS BEEN CHANGED | FROM | 12.593 | TO | 9.914 |

# APPENDIX B

## Example of Console Session Using COSMOS

```
r cosmos
W 1018.1

COSMOS READY FOR INPUT

scp data . prm
  WKM1 :
        13.0293     16.4832     11.2253     18.5775     17.4057     18.0876     34.9500
        16.4832     22.2940     14.6929     24.4822     22.8742     23.8069     45.5000
        11.2253     14.6929     10.1581     16.4694     15.4468     16.0461     30.7500
        18.5775     24.4822     16.4694     27.8051     25.8375     26.9622     51.4100
        17.4057     22.8742     15.4468     25.8375     24.5381     25.3264     48.1700
        18.0876     23.8069     16.0461     26.9622     25.3264     26.2882     50.0600
        34.9500     45.5000     30.7500     51.4100     48.1700     50.0600     98.0000

        $
swp 7 . prm . stl all 1 2 3 4 5 6 . stl ind 1 2 3 . stl dep 4 5 6 .
  WKM1 :
          .5650       .2564       .2588       .2430       .2267       .2346      -.3566
          .2564      1.1690       .4161       .6133       .5096       .5648      -.4643
          .2588       .4161       .5095       .3382       .3322       .3385      -.3138
          .2430       .6133       .3382       .8358       .5679       .7011      -.5246
          .2267       .5096       .3322       .5679       .8611       .7204      -.4915
          .2346       .5648       .3385       .7011       .7204       .7167      -.5108
          .3566       .4643       .3138       .5246       .4915       .5108       .0102

        $
save cov . std all . prm all , all .
  WKM1 :
         1.0000       .3155       .4824       .3536       .3250       .3686
          .3155      1.0000       .5392       .6204       .5079       .6170
          .4824       .5392      1.0000       .5183       .5016       .5601
          .3536       .6204       .5183      1.0000       .6694       .9059
          .3250       .5079       .5016       .6694      1.0000       .9170
          .3686       .6170       .5601       .9059       .9170      1.0000

        $
swp cov ind . prm 7 all , 7 all . std dep . prm dep , dep .
  WKM1 :
          .3358      -.4313      -.2275      -.2109       .2089       .2078       .2098
         -.4313      2.3200      -.1259     -1.0758       .1227       .1044       .1089
         -.2275      -.1259      1.2129      -.9266       .3998       .2816       .3418
         -.2109     -1.0758      -.9266      3.2660       .2749       .3690       .3299
         -.2089      -.1227      -.3998      -.2749       .4678       .2450       .3535
         -.2078      -.1044      -.2816      -.3690       .2450       .5713       .4119
         -.2098      -.1089      -.3418      -.3299       .3535       .4119       .3865

        $
  WKM1 :
         1.0000       .4739       .8313
          .4739      1.0000       .8766
          .8313       .8766      1.0000

        $
```

```
swp cov 1 . prm 7 1 dep , 7 1 dep .
  WKM1 :
          .2353      -.6312       .3712       .3484       .3628
         -.6312      1.7699       .4301       .4013       .4152
         -.3712      -.4301       .7313       .4704       .6002
         -.3484      -.4013       .4704       .7701       .6263
         -.3628      -.4152       .6002       .6263       .6193

        $
swp 2 . prm 7 1 2 dep , 7 1 2 dep .
  WKM1 :
          .3222      -.5008      -.2873       .2267       .2316       .2311
         -.5008      1.9656      -.4311       .2133       .2259       .2176
         -.2873      -.4311       .9500       .4778       .3863       .4354
         -.2267      -.2133      -.4778       .4910       .2761       .3812
         -.2316      -.2259      -.3863       .2761       .6130       .4492
         -.2311      -.2176      -.4354       .3812       .4492       .4198

        $
swp 1 . prm 7 2 dep , 7 2 dep .
  WKM1 :
          .1946      -.3972       .2810       .2892       .2865
         -.3972       .8554       .5246       .4359       .4831
         -.2810      -.5246       .5141       .3006       .4049
         -.2892      -.4359       .3006       .6390       .4742
         -.2865      -.4831       .4049       .4742       .4439

        $
swp 3 . prm 7 2 3 dep , 7 2 3 dep .
  WKM1 :
          .2556      -.2509      -.4109       .2317       .2272       .2300
         -.2509      1.2060      -.9850       .4065       .2873       .3477
         -.4109      -.9850      2.7671       .3318       .4174       .3804
         -.2317      -.4065      -.3318       .4743       .2505       .3592
         -.2272      -.2873      -.4174       .2505       .5760       .4168
         -.2300      -.3477      -.3804       .3592       .4168       .3916

        $
swp 1 . prm 7 ind ,"dep , 7 ind dep .
  WKM1 :
          .3358      -.4313      -.2275      -.2109       .2089       .2078       .2098
         -.4313      2.3200      -.1259     -1.0758       .1227       .1044       .1089
         -.2275      -.1259      1.2129      -.9266       .3998       .2816       .3418
         -.2109     -1.0758      -.9266      3.2660       .2749       .3690       .3299
         -.2089      -.1227      -.3998      -.2749       .4678       .2450       .3535
         -.2078      -.1044      -.2816      -.3690       .2450       .5713       .4119
         -.2098      -.1089      -.3418      -.3299       .3535       .4119       .3865

        $
swp 2 . prm 7 1 3 dep , 7 1 3 dep .
  WKM1 :
          .2932      -.4550      -.3847       .2839       .2606       .2739
         -.4550      2.3069     -1.1720       .1642       .1336       .1444
         -.3847     -1.1720      2.5581       .5803       .5842       .5910
         -.2839      -.1642      -.5803       .5996       .3379       .4662
         -.2606      -.1336      -.5842       .3379       .6367       .4913
         -.2739      -.1444      -.5910       .4662       .4913       .4828

        $
```

```
swp 1 . prm 7 3 dep , 7 3 dep .
  WKM1 :
          .2034      -.6158       .3163       .2869       .3024
         -.6158      1.9627       .6638       .6521       .6644
         -.3163      -.6638       .6113       .3474       .4764
         -.2869      -.6521       .3474       .6444       .4997
         -.3024      -.6644       .4764       .4997       .4918

        $
macro std cov all . prm all , all . swp cov ind . prm 7 all , 7 all .
std dep . prm dep , dep . regr .m
regr .
  WKM1 :
         1.0000       .3155       .4824       .3536       .3250       .3686
          .3155      1.0000       .5392       .6204       .5079       .6170
          .4824       .5392      1.0000       .5183       .5016       .5601
          .3536       .6204       .5183      1.0000       .6694       .9059
          .3250       .5079       .5016       .6694      1.0000       .9170
          .3686       .6170       .5601       .9059       .9170      1.0000

        $
  WKM1 :
          .3358      -.4313      -.2275      -.2109       .2089       .2078       .2098
         -.4313      2.3200      -.1259     -1.0758       .1227       .1044       .1089
         -.2275      -.1259      1.2129      -.9266       .3998       .2816       .3418
         -.2109     -1.0758      -.9266      3.2660       .2749       .3690       .3299
         -.2089      -.1227      -.3998      -.2749       .4678       .2450       .3535
         -.2078      -.1044      -.2816      -.3690       .2450       .5713       .4119
         -.2098      -.1089      -.3418      -.3299       .3535       .4119       .3865

        $
  WKM1 :
         1.0000       .4739       .8313
          .4739      1.0000       .8766
          .8313       .8766      1.0000

        $
quit .
  UDIC :
    WKPL
    WKM1
    REGR
     COV
     DEP
     IND
     ALL
    DATA
  CRPROD
        $
MEMORY BOUND 56533
TO REUSE PRIVATE DATA AND DEFINITIONS,
  START
OR
  SAVE   YRNAME
  RESUME  YRNAME
RETURNING TO CTSS.
R 40.716+14.550

logout
W 1040.8
 C0011  9900 LOGGED OUT 10/17/66    1040.9 FROM 20000.
 TOTAL  TIME USED=   1.2 MIN.
-
```

IBM